Case study

次世代シーケンサーと連携する研究インフラで生物多様性の謎にゲノムの視点から迫る

膨大な読み取りデータから遺伝子の塩基配列を再構成し、その機能を探るべくHP ProLiant DL980 G7とHP IOアクセラレータの高い処理性能をフル活用

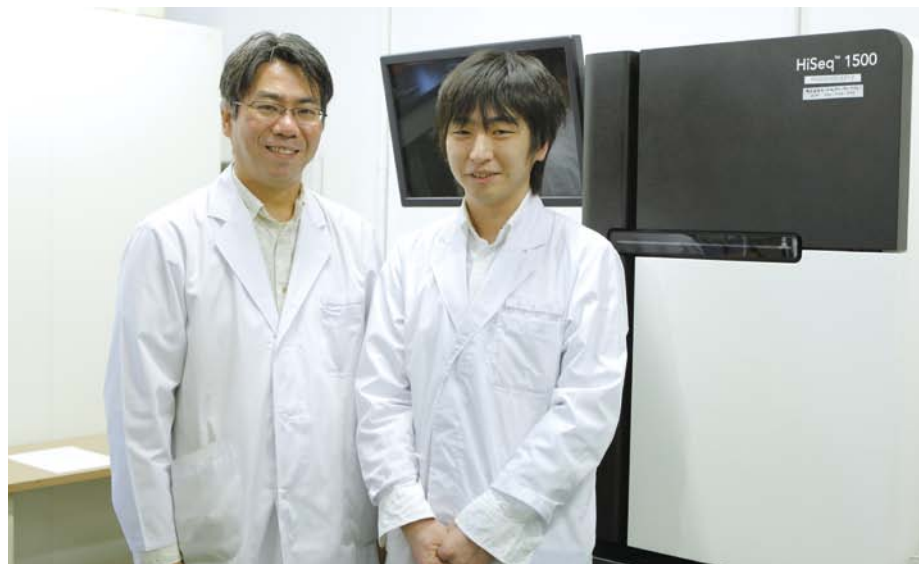
業界
研究機関

目的
次世代シーケンサーで読み取った膨大な塩基配列データの高速処理を実現する研究インフラの構築

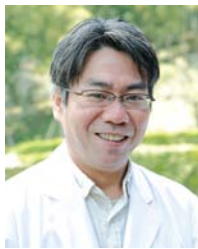
アプローチ
- 高速演算を並列化と高性能CPUで実現する
- 膨大な塩基配列データなどを展開するためにメモリー量を最大限確保する
- メモリー空間に収まりきれないデータを退避させておく高速I/Oストレージを活用する
- サーバーには高性能なインテル® Xeon® プロセッサー E7ファミリーを8基（80コア）搭載したHP ProLiant DL980 G7を採用
- ストレージには超高速I/OのHP IOアクセラレータを採用

ITの効果
- 1週間かかっていた処理がわずか2日で完了
- CPUパワー不足回避のために行っていた処理フローを工夫する手間が不要に

ビジネスの効果
- ゲノム解析の研究スピードが大幅アップ
- 遺伝子配列が未解明な生物種でも容易なゲノム解析が可能に
- これまでにない斬新なゲノム研究に道を拓く

遺伝子解析システム

導入した次世代シーケンサーを1回稼働させると、約80億塩基対分の配列データが読み取れます。遺伝子断片の数で約4000万本分、データ量でいえば数十GB。人の目どころか汎用PCですら到底手に負えない量が出力されます。理想的には、読み取り操作数回分の配列データをまとめてすべてメモリー上に展開できれば、重ね合わせなどの処理を極めて高速に行えます。メモリー量はあればあるほど解析効率も向上します。

白石 慧 氏 公益財団法人サントリー生命科学財団 生物有機科学研究所 統合生体分子機能研究部 特別研究員

終戦の翌年に早くも設立され、生命科学の最前線を拡大させる挑戦を長年にわたってリードするサントリー生命科学財団。同財団の広範な取り組みの一つとして、自らの研究機関である生物有機科学研究所が多くのユニークな基礎研究を進めている。2012年、同研究所の統合生体分子機能研究部では、生命がその進化の中で獲得してきた生物多様性の謎に遺伝子からアプローチするため、遺伝子の塩基配列を超高速で読み取ることのできる次世代シーケンサーを導入。併せて、得られた膨大な塩基配列データの分析処理を担う新たな研究インフラの構築にも着手した。中核を担うサーバーには、極めて高いパフォーマンスと巨大なメモリー量を搭載できることが求められた。

公益財団法人サントリー生命科学財団
生物有機科学研究所 統合生体分子機能研究部
部長 主幹研究員
佐竹 炎 氏

公益財団法人サントリー生命科学財団
生物有機科学研究所 統合生体分子機能研究部
特別研究員
白石 慧 氏

純粋に真理の探究に挑む取り組みを幅広く支援するサントリー生命科学財団

終戦を迎えた1945年、後にサントリーの会長となる佐治敬三氏は「これからの日本は学問や文化を通じて、世界の平和と繁栄に貢献していくべきである」との想いを強くする。そのための拠点として「純粋に真理の探究に情熱を燃やす秀れた研究者が寄り集まり、自由にテーマを選び、研究活動に没頭できる、ユニークな施設を作りたい」という氏の構想を基に、翌1946年2月にサントリー生命科学財団が設立された。

同財団は民間の資金を基盤としながらも、設立当初から、生命科学分野での学術的な基礎研究、関連する研究への資金助成といった公益性の高い活動を展開。現在では、自ら運営する生物有機科学研究所での研究事業、大学での学術研究を分析や解析などの面から支援する解析センター事業、大学の研究室や研究者への資金助成を行う研究奨励助成事業、生物有機科学研究所へのポスドク受け入れや大学院連携講座の開設などをとおした科学人材育成事業など、幅広い取り組みにより生命科学の最前線を拡大させる試みを積極的にリードしている。

2012年、同財団の生物有機科学研究所では、生命そのものの象徴であり、生命活動を司る存在でもある遺伝子をより深く、より正確に理解することを目指し、遺伝子の塩基配列の高度な解析が可能な最新鋭の次世代シーケンサーを導入。これに伴い、装置で読み取った膨大な量の塩基配列情報をバイオインフォマティクスの手法で活用するための新たな研究インフラを構築することにした。その中核として採用されたのが、高性能なインテル® Xeon® プロセッサー E7ファミリーを8基（80コア）搭載したマルチプロセッサーサーバーであるHP ProLiant DL980 G7、および超高速I/Oを実現するHP PCIe IOアクセラレータ for ProLiantサーバー（以下、HP IOアクセラレータ）であった。

次世代シーケンサーの導入に合わせ高性能な処理システムも整備する

生物有機科学研究所は、「天然有機化合物の生物活性メカニズムの解明」と「生物種の多様性と共存の神髄への肉薄」という大きく2つの研究テーマを設定。それぞれのテーマに対応して構造生命科学研究部、統合生体分子機能研究部という2つの研究室が最先端の研究活動を展開している。このうち、統合生体分子機能研究部が次世代シーケンサーの導入を決めた。

生物は長い進化の中で膨大な数の枝分かれを繰り返しながら、動物、植物、細菌と多様な生存戦略に基づき、地球上に生息域を広げてきた。そして現在では、既知の生物で約175万種、まだ発見されていないものも含めるとおよそ500万～3000万種ともいわれる多種多様な生物がそれぞれに生命活動を営み、互いに影響を及ぼし合いながら共存している。

統合生体分子機能研究部では、こうした生物多様性の巧妙なメカニズムがどのようにして生まれ、機能しているのかという疑問に対し、遺伝子や分子のレベルから迫ろうとしている。そのために、動物や植物、微生物などの幅広い生物を研究対象に、分子生物学や細胞生物学、生理学、ゲノミクスといった広範な研究手法を駆使しながら、研究を進めている。

同研究部の佐竹 炎 部長は、次世代シーケンサーを導入した狙いを次のように語る。「多様な生物それぞれの生存戦略を明らかにするには、単一の遺伝子に着目した研究だけでは不十分で、膨大で複雑な遺伝子の総体（ゲノム）を見るという、ゲノムワイドな解析を迅速に進めることが不可欠です。また、ゲノムの解析は、その配列が既知の生物種だけでなく、時にはまだきちんとしたリファレンスのない生物種に対しても行う必要があります。こうした取り組みを進めるには、より高性能でより高速な読み取りが可能な次世代シーケンサーの導入を決めました」。

「しかし」と佐竹部長は続ける。「次世代シーケンサーの能力を効果的に活用するには、これまでは想像もできなかった膨大な量の読み取りデータを、バイオインフォマティクスの手法で極めて高速に処理できる解析システムが不可欠です。研究部内で共通に使える研究インフラとしても、早急に整備する必要がありました」。

研究インフラのハードで重視したのは搭載メモリー量の大きさと高速I/O

生物の細胞内には、生命活動を支えるタンパク質の合成にかかわるDNAやRNAなどの遺伝子が数万種類存在する。これらの遺伝子はおよそ1万塩基ほどの長さを持つ。生物多様性の秘密を探っていくには、これらの遺伝子の塩基配列を解明することが最初のステップになる。

導入した次世代シーケンサーと新たに整備する研究インフラを組み合わせ、遺伝子の塩基配列を解析していくプロセスは次のようなものだ。まず、試料となる細胞から、塩基配列を読み取りたいDNAやRNAを抽出する。次に、これらを、次世代シーケンサーの読み取り性能に合わせて、数百～数千塩基ほどの短い鎖に断片化。塩基配列を次世代シーケンサーでひたすら自動的に読んでいく。得られた塩基配列データは研究インフラに取り込み、重ね合わせなどの処理を繰り返しながら全体の塩基配列を推定、断片

■研究インフラの構成図

インテル® Xeon® プロセッサー
E7ファミリー

ハードウェア:HP ProLiant DL980 G7
搭載プロセッサー:インテル® Xeon® プロセッサー
E7-4870 2.4GHz 8基／80コア
搭載メモリー:1TB
オプション:HP PCIe IOアクセラレータ for ProLiantサーバー 1.2TB×2台
OS:Red Hat Enterprise Linux 6.2

化する前の塩基配列を再構成する。

「導入した次世代シーケンサーを1回稼働させると、約80億塩基対分の配列データが読み取れます。断片の数に換算すると約4000万本。データ量でいえば数十GBという、人の目どころか汎用PCですら到底手に負えない量が出力されます。理想的には、読み取り操作数回分の配列データをまとめてすべてメモリー上に展開できれば、重ね合わせなどの処理を極めて高速に行えます。また、得られた配列データには、装置の特性上、読み取りミスの発生する可能性があります。遺伝子塩基配列がすでに分かっている生物種であればあまり影響はないのですが、未解明の生物種の場合には推定の精度を上げるために読み取りミスを許容して処理しなくてはなりません。候補の塩基配列が増える分、必要なメモリー量は増大。メモリー量はあればあるほど解析効率も向上します」。新しい研究インフラの整備を担当した同研究所特別研究員の白石 慧 氏は、中核となるサーバーを探すに当たって、最も重視したポイントをこう解説する。

さらに、サーバーと接続するストレージの性能にも白石氏は注目していた。「読み取った塩基データは膨大になるため、すべてをメモリー上に展開しておくことは現実的に困難です。もちろん、重ね合わせ処理を行うためのメモリー領域も残しておく必要もあります。ある程度の塩基データは外部ストレージに退避させておき、必要となったときに高速にメモリー上へ展開する。こうした使い方になることが想定されたため、高速I/Oの可能なストレージが必要になるだろうと考えていました」(白石氏)。

京都大学での検証実績を見て HP製ハードウェアの可能性を確信

具体的な機種選定にあたり白石氏が参考にしたのは、自身が以前所属していた京都大学大学院薬学研究科の奥野恭史教授が率いるシステム創薬科学研究室で行われた検証での実績だった。

同研究室では、体内にある遺伝子やタンパク質と化学物質との相互作用をコンピューター上で効率的に推定し創薬への応用を探るという、インシリコ創薬の研究に取り組んでいる。研究プロセスの中で大きな鍵を握るのは、10の60乗を超える天文学的な数の化学物質の中から、特定の遺伝子やタンパク質と相互作用する可能性がある化学物質を精度良く、短時間でスクリーニングできようにする新たなバーチャルスクリーニング技術だ。

この処理の検証用システムとして、80コアHP ProLiant DL980 G7とHP IOアクセラレータという構成が使用された。システムに求められた要件は、強力なCPUパワー、処理の並列化、大きな搭載メモリー量、I/O性能の高い外部ストレージ、というものであり、白石氏が探していた構成要件とまさにピタリと一致した。

「スクリーニングを行う前に、判断基準を提供する知識モデルを機械学習によって構築するのですが、入力するサンプル数が10倍になると必要なメモリー量は100倍に、100倍では1万倍に、と2乗のオーダーで計算量も必要なメモリー量も増えていきます。HP ProLiant DL980 G7とIOアクセラレータを使った検証では、知識モデル構築にかかる時間などを、入力サンプル数を従来の100倍に増やした場合で評価したようですが、十分に満足できる結果が得られたということでした。整備を任された研究インフラもこのハードウェア構成でいけるだろう、と確信できました」(白石氏)。

遺伝子の解析だけでなく、機能の解明でも大きな期待

新しい研究インフラの整備作業は2013年4月からスタート。これまでのところ、ゲノミクス領域のモデル生物として良く知られ、統合生体分子機能研究部でも長年研究してきた「ホヤ」の一種をモデル試料として選定。次世代シーケンサーと研究インフラを組み合わせて効率良く、高い再現性で断片化前の遺伝子の塩基配列を推定する処理フローを、使用するソフトウェアの選定なども含めて、確立する作業が進んでいる。

こうした遺伝子の塩基配列の「解析」というテーマに加え、同研究部では塩基配列の「解釈」という二つ目のテーマについても、HP ProLiant DL980 G7とIOアクセラレータから成る研究インフラを活用したいと考えている。「解釈」とは、判明した遺伝子塩基配列のうち、ある部分がどのような機能を果たしているかを予測するというものだ。これは白石氏がこれまで取り組んできたメインの研究テーマでもある。

「特定の塩基配列とそれが実現している機能との関係性を明らかにしていくという研究であり、これは奥野研究室で開発した独自のバーチャルスクリーニング技術を拡張し、適用することで可能だと考えています。奥野研究室の検証結果から、この解釈というテーマでも、研究インフラが大きく役立ってくれることになるでしょう」と、白石氏は期待を込める。

すでに、以前の研究で使っていたデータを持ち込み、整備中の研究インフラでテストを行ったという。「以前のハードウェア環境では結果が出るまでに1週間かかっていました。それが、HP ProLiant DL980 G7上で同じ処理を走らせたところ、わずか2日で完了したのです。以前であれば、計算に必要なパラメーターの検討だけでこの1週間単位の計算を複数行わなくてはなりませんでした。検討にかかる時間もメモリーの節約法を検討する手間も大きく省かれ、HP ProLiant DL980 G7のメリットを身をもって体感できました」(白石氏)。

「統合生体分子機能研究部では、ホヤをはじめ、マウスや植物、細菌など広範な生物種を研究対象としており、まだ遺伝子の塩基配列が解明されていない生物種も扱っています。研究員が共通に利用できる解析の処理フローを確立するだけでも、研究スピードは大幅な向上が見込めます。そして、これらの膨大なデータに対してユニークな解析を試みることで、当研究部から画期的な知見を発信できると確信しています。新しい研究インフラには大きな期待を持っています」と、佐竹部長は笑顔で話を締めくくった。

ソリューション概略

導入ハードウェア

HP ProLiant DL980 G7
HP PCIe IOアクセラレータ for ProLiantサーバー

お問い合わせはカスタマー・インフォメーションセンターへ
03-5749-8328 月〜金 9:00〜19:00　土 10:00〜17:00(日、祝祭日、年末年始および5/1を除く)
機器のお見積もりについては、代理店、または弊社営業にご相談ください。
HP ProLiantに関する情報は http://www.hp.com/jp/proliant

本カタログは、環境に配慮した用紙と植物性大豆油インキを使用しています。

日本ヒューレット・パッカード株式会社
〒136-8711 東京都江東区大島2-2-1